AN ANALYSIS OF COMPUTER SCIENCE EDUCATION PUBLICATION USING LOTKA'S LAW[*]

Christopher R. Merlo
Nassau Community College
1 Education Dr.
Garden City, NY 11530
(516) 572-7383
cmerlo@ncc.edu

Lori Hoeffner
Adelphi University
1 South Ave.
Garden City, NY 11530
(516) 877-3232
hoeffner@adelphi.edu

Joan M. Merlo
Molloy College
1000 Hempstead Ave.
Rockville Centre, NY 11570
(516) 678-5000 x6442
jmerlo@molloy.edu

Richard Moscatelli
Nassau Community College
Garden City, NY 11530
(516) 572-7383
moscatr@ncc.edu

ABSTRACT

This study helps to determine what information is shared among Computer Science educators, and how that information is shared, so that we may better prepare our students to identify emerging trends in the discipline. An analysis of publication activity through the use of bibliometric and other techniques is presented. Productivity of first authors in the field of Computer Science Education is considered. Publication data from five leading journals in Computer Science Education (*n* = 5,274 articles) were analyzed using Lotka's Law (1926). Results indicate that the productivity distribution of first authors is inconsistent with Lotka's distribution. Conclusions are presented about Computer Science Education research trends, and about future research paths.

INTRODUCTION

A significant portion of what Computer Science students learn, and how they learn it, is a result of what their teachers publish. This paper introduces bibliometric modeling in an effort to better understand the publishing patterns in Computer Science Education that make this material available to Computer Science educators and their students. The use of bibliometric modeling within a particular discipline advances the understanding of that discipline and how knowledge is organized and shared within it. This study is an important step in determining what information is shared among Computer Science educators, and how that information is shared, so that we may better prepare our students to identify emerging trends in the discipline. Findings suggest that Computer Science Education enjoys both a breadth and a depth of research and researchers, even though this study was limited to examining only first authors in an attempt to replicate Lotka's work. While it is understood that this approach to analyzing Computer Science Education literature is different, the contribution will be significant because it will help to codify the presently emerging trends, and help to predict those that will emerge in the future. In a field that changes as rapidly as Computer Science, it is critical to be able to identify leading scholars and trends. Future research will utilize co-citation analysis to discover the most significant clustering of research topics.

The method presented here to examine the literature is a two-step bibliometric process, that enables a better understanding of the most productive first authors (authors whose name appears first in the citation), and secondly, an application of Lotka's Law [1] in order to predict the likelihood that a first author who has written one article will produce more articles as first author. The decision to examine first authors, instead of all authors, was made in order to parallel Lotka's original study. Lotka examined the rate at which people contribute articles as first authors in their field, specifically those who publish multiple articles as opposed to those who publish only one [5]. His observation, since referred to as Lotka's Law, predicts an inverse exponential relationship between the amount of first authors who write multiple papers and the amount who write only one [5]. Schorr [12] and Murphy [8] attempted to determine whether Lotka's Law could be applied to non-scientific productivity. Coile [1] examined these two applications and determined that both Murphy's and Schorr's results were flawed because they ignored the inverse exponential relationship. Thus, Coile [1] reevaluated their original data by correctly applying Lotka's Law. The present study reinforces Coile's techniques and examines literature in Computer Science Education.

Other authors [6] have performed a citation analysis on Computer Science Education literature, but none have performed a first-author frequency analysis in the field. Similarly, others have tested Lotka's Law by applying a variety of techniques [1, 7, 8, 9–13], but none have attempted to apply Lotka to Computer Science Education literature. Some have changed Lotka's rubric by including all authors, not simply first authors [11]; others have attempted new measures for predicting scientific productivity, namely rate of publication and career duration of authors [3,4]. One study only extracted data from a single journal [8]. The present paper contributes to the field of Computer Science Education by considering authors who have published in five Computer Science Education journals.

METHOD

Citations of articles appearing in *SIGCSE Bulletin* (*n* = 2,498), *The Journal of Computing Sciences in Colleges* (*n* = 1,605), *Computer Science Education* (*n* = 299), *The Journal of Computing in Small Colleges* (*n* = 177), and *Journal on Educational Resources in Computing* (*n* = 117) from 1989 through 2008 were obtained from the ACM digital library, or, in the case of *Computer Science Education*, the publisher's web site (Taylor and Francis). These five journals were selected because each printed more than 100 articles related to Computer Science Education during the 20 year period under consideration (1989 through 2008), and because they emphasized the teaching of Computer Science in postsecondary institutions. The search was limited to peer-reviewed articles and excluded book reviews, editorials, announcements, abstracts, panels, posters, tutorials, workshops, and working groups. Variables extracted from the citations included year of publication, author name, title of article, and title of publication. PHP programs were written to extract these data from article citations, using regular expression matching, and to store author and article data in a MySQL database. These data could then be counted, and expected first-author productivity levels could be calculated.

Many first authors who wrote multiple articles were cited with different spellings of their name, or with and without their middle initial, or with other inconsistencies. For example, Ellen Walker is listed as "Ellen Walker" with three citations, and also as "Ellen L. Walker" with five citations; Renée McCauley is listed with five different spellings, and so on. Verification was accomplished by a comparison of institutional affiliation to insure the individuals with inconsistent name formatting were actually the same person. In all, there were 672 names that were corrected, so that each person was listed correctly with the appropriate number of citations. This recoding of multiple authors with variation in name format was important to insure that all authors received the correct number of citations, a critical step necessary to insure appropriate application of Lotka's Law.

Table 1: Amount of First Authors, 1989-2008

Amount of First Authors	Number	Percent
First Authors Who Wrote Only One Article	2,083	69.27%
First Authors Who Wrote More Than One Article	924	30.73%
Total	**3,007**	**100%**

Finally, 5,274 articles were selected from the five journals over the twenty year period under investigation. From these 5,274 articles, 3,007 first authors were identified. (See Table 1.) First authors are those authors whose name appears first in an article's citation. Some authors wrote more than one article; therefore, the total number of first authors (3,007) is fewer than the total number of articles (5,274). These data provide interesting insights into author productivity within the field of Computer Science Education. As Table 2 indicates, the greatest number of articles was written by David Ginat ($n = 23$); 30 other first authors wrote as many as 10 articles, and 2,083 (69%) of the 3,007 first authors wrote only one article. These data provide the basis for the application of Lotka's Law.

RESULTS

In 1926 Alfred J. Lotka attempted to determine "...the part which men of different calibre contribute to the progress of science" by counting the amount of first authors in two scientific journals, which he considered separately from each other, and the amount of papers each author published. He found that the relationship between the frequency f of persons making c contributions is: $c^n f =$ const.

Lotka determined that the exponent n was approximately equal to 2 in both of his data sets, which means that "...the number [of first authors] making n contributions is about $1/n^2$ of those making one." [5]. Moreover, "...according to the inverse square law, the proportion of all contributors who contribute a single [article] should be just over 60 per cent [60.79%]."[5]. Therefore, the exponent of 2 was used in the present study to compute the expected amounts and percentages of first authors who contributed one article, two articles, etc. Using the expected percentages f_c of contributors making c contributions, $F_0(X)$ was calculated, which is the cumulative value of f_c for the values of c in the range $1 \leq c \leq 23$. $S_c(X)$, the cumulative value of observed percentages of contributors making c contributions, was also calculated. (See Table 3, "First Application: $n = 2$.") The Kolmogorov-Smirnov (K-S) statistic at the 0.01 level of significance is calculated by dividing 1.63 by the square root of the amount of observed first authors (3,007) [1]; the K-S statistic for the present study

Table 2: First Authors Producing Ten or More Articles Published in *SIGCSE Bulletin*, *The Journal of Computing Sciences in Colleges*, *Computer Science Education*, *Journal of Computing Sciences in Small Colleges*, and *Journal on Educational Resources in Computing*, 1989-2008

First Author	Occurrences	First Author	Occurrences
Ginat, David	23	Armoni, Michal	11
Kumar, Amruth	21	Cunningham, Steve	11
Roberts, Eric	20	Maurer, Ward Douglas	11
Astrachan, Owen	18	Proulx, Viera	11
Cassel, Lillian	16	Cliburn, Daniel C.	10
Ben-Ari, Mordechai (Moti)	15	Kurtz, Barry L.	10
Sanders, Dean	15	Parlante, Nick	10
Bergin, Joseph	14	Pheatt, Charles	10
Becker, Katrin	13	Rößling, Guido	10
Gal-Ezer, Judith	13	Rasala, Richard	10
Naps, Thomas	13	Robbins, Steven	10
Adams, Joel	12	Rodger, Susan	10
Impagliazzo, John	12	Scott, Terry	10
Wick, Michael	12	Soh, Leen-Kiat	10
Wolz, Ursula	12	Tucker, Allen	10
Almstrum, Vicki	11		

equals 1.63 ÷ √3007 = 0.0297. The K-S test is a comparison of the K-S statistic and the maximum deviation of $F_0(X)$ and $S_c(X)$, denoted D as: $D = \max|F_0(X) - S_c(X)|$. The maximum deviation D must be smaller than the K-S statistic in order for the K-S test to indicate conformity to Lotka's Law. Applying the K-S test to the present study, using $n = 2$ as the exponent in the equation $c^n f = \text{const}$, the value of D is 0.0848. This value exceeds the K-S statistic (0.0848 > 0.0297), indicating that the observed data do not conform to Lotka's Law at the 0.01 level of significance. (See Table 3, "First Application: $n = 2$.").

In his analysis of first author productivity in Auerbach's *Geschichtstafeln der Physik*, Lotka plotted the frequencies of first authors who had contributed 1, 2, 3, etc. papers against these numbers 1, 2, 3, etc., on logarithmic axes. Lotka then used least-squares regression analysis to determine the slope of the best-fitting line to these data points, which was 2.021. Because this value is so close to 2, Lotka chose to use $n = 2$ as the exponent in the equation $c^n f = \text{const}$ when calculating his expected amounts and percentages of first-author productivity. However, when he performed the same calculation on the data he observed in the decennial index of *Chemical Abstracts* 1907-1916, the slope of the best-fitting line was $n = 1.888$, too far removed from 2 to force these data into the inverse-square law. Lotka chose instead to base his expected amounts and frequencies of first-author productivity in *Chemical Abstracts* on this calculated exponent.

Following Lotka, data from the present study were also plotted on logarithmic axes, and the slope of the best-fitting line was found to be 2.713. Because of the magnitude of the difference between 2 and 2.713, a decision was made to recalculate the expected percentages f_c and re-apply the K-S test, using this slope 2.713 as the exponent n in the equation $c^n f = \text{const}$. This allowed for a second application of the K-S test to the present data. The cumulative percentage $F_0(X)$ was recalculated. As a result, the maximum deviation D of $F_0(X)$ and $S_c(X)$ was determined to be 0.0959; therefore, since D exceeded the K-S statistic (0.0959 > 0.0297), the K-S test again failed to indicate conformity. (See Table 3, "Second Application: $n = 2.713$.")

CONCLUSION

These results suggest that the productivity distribution of first authors in Computer Science Education does not follow Lotka's Law. Instead, Computer Science Education researchers communicate with each other, and therefore with their students, through published works in ways that are unexpected. This conclusion will impact the kind of references that faculty make in the classroom. The percent of first authors who wrote only one article in *SIGCSE Bulletin*, *The Journal of Computing Sciences in Colleges*, *Computer Science Education*, *The Journal of Computing in Small Colleges*, and *Journal on Educational Resources in Computing* during the timeframe of this study is 69.27%. This is significantly different than one would anticipate given Lotka using the ideal exponent of $n = 2$ (60.79%) [2], or the computed exponent of $n = 2.713$ (78.86%). On the other hand, only 20 of the 3,007 authors in the present study – two thirds of one percent – wrote more than 10 papers as first author during the 20 years of the study, while approximately 7 out of 10 (2,083) wrote only one as first author. It seems that the research fronts in the field of Computer Science Education are being led by a small, very prolific group of first authors. At the same time, thousands of new first authors contributed to their field, adding to the body of knowledge available for classroom discussions and assignments. This suggests that Computer Science Education enjoys both breadth and depth of research and researchers. These kinds of contributions, although they do not conform to Lotka's Law, help maintain research fronts within the profession, provide novel approaches to teaching Computer Science, and foster communication among Computer Science professionals. Future research will involve an analysis of Lotka's law that considers all authorship equally (not just first authorship) in order to provide an alternative model that more accurately reflects Computer Science Education literature. This will help to identify whose new ideas and techniques are being used in Computer Science classrooms, and potentially whose ideas deserve more attention from CS educators.

Table 3: Two Applications of the Kolmogorov-Smirnov Test for Goodness-of-Fit of Observed First Author Productivity to Lotka's Law

c	Expected % of authors who wrote *c* articles	Expected amount of authors who wrote *c* articles	Observed amount of authors who wrote *c* articles	Observed % of authors who wrote *c* articles	Observed amount as % of expected amount	$F_0(X)$ (Cumulative expected %)	$S_c(X)$ (Cumulative observed %)	$\|F_0(X) - S_c(X)\|$
First Application: Using Exponent *n* = 2								
1	60.79%	1828.04	2083	69.27%	113.95%	60.79%	69.27%	0.0848
2	15.20%	457.01	444	14.77%	97.15%	75.99%	84.04%	0.0805
3	6.75%	203.12	195	6.48%	96.00%	82.75%	90.52%	0.0778
4	3.80%	114.25	103	3.48%	90.15%	86.55%	93.95%	0.0740
5	2.43%	73.12	62	2.06%	84.79%	88.96%	96.01%	0.0703
6	1.69%	50.78	39	1.30%	76.80%	90.67%	97.31%	0.0664
7	1.24%	37.31	19	0.63%	50.93%	91.91%	97.94%	0.0603
8	0.95%	28.56	17	0.57%	59.52%	92.86%	98.50%	0.0565
9	0.75%	22.57	14	0.47%	62.03%	93.61%	98.97%	0.0536
10	0.61%	18.28	11	0.37%	60.17%	94.21%	99.33%	0.0512
11	0.50%	15.11	5	0.17%	33.10%	98.72%	99.50%	0.0478
12	0.42%	12.69	4	0.13%	31.51%	95.14%	99.63%	0.0450
13	0.36%	10.82	3	0.10%	27.73%	95.50%	99.73%	0.0424
14	0.31%	9.33	1	0.03%	10.72%	95.81%	99.77%	0.0396
15	0.27%	8.12	2	0.07%	24.62%	96.08%	99.83%	0.0375
16	0.24%	7.14	1	0.03%	14.00%	96.32%	99.87%	0.0355
17	0.21%	6.33	0	0.00%	0.00%	96.53%	99.87%	0.0334
18	0.19%	5.64	1	0.03%	17.72%	96.71%	99.90%	0.0319
19	0.17%	5.06	0	0.00%	0.00%	96.88%	99.90%	0.0302
20	0.15%	4.57	1	0.03%	21.88%	97.04%	99.93%	0.0290
21	0.14%	4.15	1	0.03%	24.12%	97.17%	99.97%	0.0279
22	0.13%	3.78	0	0.00%	0.00%	97.30%	99.97%	0.0267
23	0.11%	3.46	1	0.03%	28.94%	27.41%	100.00%	0.0259
Second Application: Using Exponent *n* = 2.713								
1	78.86%	2371.30	2083	69.27%	87.84%	78.86%	69.27%	0.0959
2	12.03%	361.65	444	14.77%	122.77%	90.89%	84.04%	0.0685
3	4.00%	120.38	195	6.48%	161.99%	94.89%	90.52%	0.0437
4	1.83%	55.16	103	3.43%	186.74%	96.72%	93.95%	0.0278
5	1.00%	30.11	62	2.06%	205.93%	97.73%	96.01%	0.0172

6	0.61%	18.36	39	1.30%	212.42%	98.34%	97.31%	0.0103
7	0.40%	12.08	19	0.63%	157.22%	98.74%	97.94%	0.0080
8	0.28%	8.41	17	0.57%	202.09%	99.02%	98.50%	0.0051
9	0.20%	6.11	14	0.47%	229.09%	99.22%	98.97%	0.0025
10	0.15%	4.59	11	0.37%	239.56%	99.37%	99.33%	0.0004
11	0.12%	3.55	5	0.17%	141.02%	99.49%	99.50%	0.0001
12	0.09%	2.80	4	0.13%	142.85%	99.58%	99.63%	0.0005
13	0.07%	2.25	3	0.10%	133.13%	99.66%	99.73%	0.0007
14	0.06%	1.84	1	0.03%	54.26%	99.72%	99.77%	0.0005
15	0.05%	1.53	2	0.07%	130.85%	99.77%	99.83%	0.0006
16	0.04%	1.28	1	0.03%	77.95%	99.81%	99.87%	0.0005
17	0.04%	1.09	0	0.00%	0.00%	99.85%	99.87%	0.0002
18	0.03%	0.93	1	0.03%	107.29%	99.88%	99.90%	0.0002
19	0.03%	0.80	0	0.00%	0.00%	99.91%	99.90%	0.0001
20	0.02%	0.70	1	0.03%	142.79%	99.93%	99.93%	0.0000
21	0.02%	0.61	1	0.03%	163.00%	99.95%	99.97%	0.0001
22	0.02%	0.54	0	0.00%	0.00%	99.97%	99.97%	0.0000
23	0.02%	0.48	1	0.03%	208.63%	99.99%	100.00%	0.0001

.

REFERENCES

[1] R. C. Coile, Lotka's frequency distribution of scientific productivity, *Journal of the American Society for Information Science*, 366–370, 1977.

[2] L. Egghe, Relations between the continuous and the discrete Lotka power function, *Journal of the American Society for Information Science and Technology*, 664–668, 2005.

[3] J. C. Huber, A new model that generates Lotka's law, *Journal of the American Society for Information Science and Technology*, 53, (3), 209–219, 2002.

[4] J. C. Huber and R. Wagner-Dobler, Scientific production: A statistical analysis of authors in physics 1800-1900, *Scientometrics,* 50, 437–453, 2001.

[5] A. J. Lotka, The frequency distribution of scientific productivity, *Journal of the Washington Academy of Sciences*, 16, (12), 317–323, 1926.

[6] R. Lister and I. Box, A citation analysis of the SIGCSE 2007 proceedings, *SIGSCE Bullitin,*, 476-480, 2008.

[7] J. M. Merlo, L. Hoeffner, and C. R. Merlo, An analysis of institutional research publications in higher education: The communication of knowledge, paper presented at the Annual Forum of the Association for Institutional Research, Seattle, Washington, 2008.

[8] L. J. Murphy, Lotka's law in the humanities?, *Journal of the American Society for Information Science*, 461–462, 1973.

[9] M. L. Pao, An empirical examination of Lotka's law, *Journal of the American Society for Information Science*, 37, (1), 26–33, 1985.

[10] W. G. Potter, Lotka's law revisited, *Library Trends*, 30, (1), 21–39, 1981.

[11] T. Radhakrishman and R. Kernizan, Lotka's law and computer science

literature, *Journal of the American Society for Information Science*, 51–54, 1979.

[12] A. E. Schorr, Lotka's law and map librarianship, *Journal of the American Society for Information Science*, 189–190, 1975.

[13] H. Voos, Lotka and information science, *Journal of the American Society for Information Science,* 25, (4), 270–272, 1974.